ANNUAL PROGRESS REPORT

BROWSER
An Automatic Indexing On-Line Text Retrieval System

CONTRACT NONR 4456(00)

Submitted to:

Information Systems Branch
Office of Naval Research
Department of the Navy
Washington, D. C. 20360

Prepared by:

J. H. Williams, Jr.

September 1969

Federal Systems Division
INTERNATIONAL BUSINESS MACHINES CORPORATION
Gaithersburg, Maryland 20760

INTRODUCTION

Text retrieval systems all have the same goal yet employ different techniques which affect facility of use and effectiveness of retrieval. It is easy to visualize that a machine can match identical terms of a query with those of the search file. However, the crucial problem lies in obtaining those relevant documents which do not contain the identical set of terms that the searcher thought they would. Until algorithms are developed that will recast the searcher's problem in as many different expressions that author's have employed, the machine must remain lowly and leave the thinking to the searcher.

The design of text retrieval systems center around the minimization of two types of errors: missing true hits, and retrieving false hits. As in other human and physical systems the errors are interrelated since a decrease in one will cause an increase in the other. Many compensation techniques have been offered to reduce either of the errors separately. A Boolean statement query attempts to decrease the false hits but due to its preciseness causes an increase in missed documents. An internal table of related terms (thesaurus) attempts to decrease the misses but due to the broadening of the query increases the false hits.

What is needed is a system of techniques under the control of the user with which he may reduce the mismatch between his familiarity

of word usage in the desired subject area and word usage employed in the documents to be searched. In addition, these techniques must not place an additional burden on the searcher, but should appeal to him naturally whether he is just entering the field or has considerable familiarity with this field.

Keyword indexing and Boolean statement queries place additional barriers between the user and the desired document. Keyword dictionaries create a specific vocabulary that must be learned. A Boolean searching system is biased toward minimizing the false hits at the expense of not retrieving documents that are "close" or those that just miss the rigid logical conditions of the query. In order to achieve preciseness, Boolean search systems restrict the searcher's natural mode of expression by requiring the query to be expressed as a logical condition.

An alternative to the Boolean searching strategy has been developed, programmed and tested.* This system, called BROWSER, (BRowsing On-line With SElective Retrieval), allows the searcher to express a query in his own words, in an unrestricted form. The system provides many points of interaction through the IBM 2260 display terminal.

Rather than biasing the system toward minimizing the false hits, the basic design is biased toward minimizing the number of misses. Also provided are user controls with which he can broaden or narrow the effect of his query, depending upon his daily needs.

*Developed under joint support by Office of Naval Research and the International Business Machines Corporation.

The assumption underlying the BROWSER system is that users feel more comfortable when more documents than required are delivered. Then the user may screen out those not meeting his precise requirement and while doing so he also sees documents that are related in many directions to his primary concern. Thus, for those searches in which there is no direct response, he knows that he has reviewed the surrounding material. For example, a successful patent search is one in which no relevant documents are found; however, the searcher must review documents which are "close" to the new invention.

This paper describes a text retrieval system (BROWSER) which emphasizes the retrieval by man and machine of relevant documents containing terms that the searcher had not anticipated. The first section lists the requirements for which the system was designed, and the system features resulting from those requirements. The next section describes the BROWSER system and provides a search example. The last section describes an evaluation of the system based on more than 100 queries.

REQUIREMENTS

The IBM International Patent Operations Department, after surveying existing narrative text searching systems and finding that they did not meet their requirements, requested the Federal Systems Division to develop and test a system that would meet the following requirements:

1. The system should be capable of automatically indexing abstracts.

2. A patent attorney should be able to formulate his query in language he would normally use in writing to another patent attorney, without the need of a query specialist interposed between him and the system.

3. The system should be capable of indexing and retrieving abstracts written in English, French or German.

4. The system should be capable of satisfying different degress of specificity and exhaustiveness.

5. The system should rank abstracts in decreasing likelihood of relevancy so that abstracts directly matching a portion of the query may be reviewed first.

6. The design approach should provide effective retrieval from data bases consisting of hundreds of thousands of abstracts.

7. The system should provide for searching the structured bibliographic fields such as date, inventor, and company; these in conjunction with a text search.

8. The system should provide for on-line terminals with browsing capability.

SYSTEM DESCRIPTION

The three major steps of the system are: creation of the dictionary, automatic indexing and creation of a search file, and executing a search.

Creation of Dictionary

The search term dictionary is created by inputting the title and text of each abstract to a word scan program, which identifies each word and creates a file of all text words except the common function words, such as "the," "for," "with," etc. The text words are sorted in alphabetical sequence and listed along with a count of the number of abstracts in which each word occurred.

The list of words is reviewed in order to determine a single form of a word to represent a class of words. The set of these single forms constitute the search term dictionary and become the entries to the search file. For example, the character string TRAIN is the dictionary entry representing TRAIN, TRAINS, TRAINED, and TRAINING.

Automatic Indexing and Creation of Search Files

The abstracts are automatically indexed by matching each text word in each abstract (except function words) with each search term. The match condition is met when the longest search term has been found which matches all or a portion of an input word. On each

occurrence of such a longest match, the abstract number is recorded in the search file. This form of a full text search file is referred to as an inverted file index. The inverted file is stored on a direct access device, which enables each search term document string to be retrieved independently. After the inverted file is created, a weighted value is computed for each term. Since the weighting value is a function of the number of abstracts in the file, it will continually reflect the effect of new abstracts added to the file.

Another file is also stored on a direct access device for display and printing functions. This file consists of the full text of each abstract in its original form, along with its title and structured bibliographic information.

A third file is created for applications requiring searches of the structured bibliographic information, containing the necessary formatted fields.

Executing a Search

A search can be executed either in off-line (batch) mode or on-line mode with a display terminal. A search is initiated by preparing a query written in natural language. Each term of the query is matched with the search term dictionary. Results of this matching yields a list of search terms representing the query. The inverted file is then accessed for each query search term and an information value is computed for each

abstract. The abstracts are then sorted in descending information value sequence.

In addition to the traditional outputs of citations and the abstract text, a new form of output is provided by this system. The new output, called the response index, is an index to the search results, similar to a KWIC index. The response index consists of a one line entry for each abstract retrieved. The entry contains each of the search terms (in abbreviated form) occurring in that abstract. The sequence of the index can be based either on the ranking of the abstracts or on the weights of individual search terms.

In the off-line mode, the abstract text and response index are printed at the end of the search. In the on-line mode they are printed on demand. In the on-line mode, the response index is presented on the IBM 2260 display terminal. The searcher can display successive abstracts in their ranked sequence, review them, and selectively print those meeting his relevance requirement. The searcher can override the ranked sequence by selecting a subset of query terms which define a meaningful concept to him. A response index consisting of only these terms is displayed. The searcher now scanning the response index may display the abstract of any document containing the terms he believes represent the desired concept. Upon review, the entire abstract may be printed if it meets the searcher's relevance requirement. The searcher may reselect or rearrange the subset of query terms as often as necessary.

SEARCH EXAMPLE

A sample query will now be used to show the various points of interaction and the information fed-back to the user. The query shown in Figure 1 was input to a data base of 8000 abstracts most of which were published between 1963 and 1967 by the Defense Documentation Center or Department of Commerce Clearinghouse for Scientific and Technical Information. These abstracts were obtained from the IBM Technical Information Retrieval Center. Thus only abstracts within IBM's field of interest are in the test data base.

The query was input and each of its terms was matched with the search term dictionary. Figure 2 contains the search terms extracted, the number of documents indexed by the term, and their weighting as displayed to the searcher on the screen. The searcher reviews the search terms and observes their document frequencies. He may delete any terms having high frequencies or that are not crucial to his relevance criteria, e.g., INTEREST and INCLU. Upon completing his review, he executes the search.

Figure 3 displays the output options available after the search is completed. Two sets of output options are provided. The first set causes off-line printing; the second set displays on-line. The number of responses are also shown at this time.

I am interested in the automatic classification of documents into subject categories, groups, clusters or clumps, using discriminant or latent factors. Other statistical indexing techniques include association matrices or correlation coefficients based on word occurrences.

Figure 1. Query

ASSOC	440	4.2	AUTOMAT	479	4.1	CATEGOR	123	6.0
CLASSIF	216	5.2	CLUMP	7	10.2	CLUSTER	28	8.2
COEFF	328	4.6	CORRELAT	308	4.7	DISCRIM	77	6.7
DOCUMENT	263	4.9	FACTOR	503	4.0	GROUP	336	4.6
INDEX	268	4.9	INCLU	1376	2.6	INTEREST	336	4.6
LATENT	11	9.5	MATRI	219	5.2	OCCUR	149	5.8
STATIST	316	4.7	SUBJECT	484	4.1	TECHNIC	359	4.4
WORD	171	5.6						

Figure 2. Search Terms

OFF-LINE PRINTING

List search terms
List response index in document value sequence
List response index in term value sequence
List N abstracts
List N titles
List current abstract displayed

ON-LINE DISPLAY

Display search terms
Select browsing terms
Display response index
Display abstract
Display N titles

Figure 3. Output Options

ASSOC	GROUP	AUTOMAT
AUTOMAT	INDEX	CLASSIF
CATEGOR	LATENT	DOCUMENT
CLASSIF	MATRI	DISCRIM
CLUMP	OCCURR	STATIST
CLUSTER	STATIST	FACTOR
COEFF	SUBJECT	LATENT
CORRELAT	TECHNIC	WORD
DISCRIM	WORD	
DOCUMENT		
FACTOR		

Figure 4a. Search Terms

Figure 4b. Selected Browsing Terms

The browsing facility allows the searcher to review an index to his search result, review the text of retrieved abstracts and selectively print abstracts responding to his needs. The searcher selects a subset of search terms extracted from the query which represent his desired concept to display abstracts containing the terms. The arrangement of the displayed terms is determined by the user. The terms may be selected in any sequence which forms the most useful patterns for browsing.

Figure 4a shows the search terms representing the query, and Figure 4b shows the sequence of browsing terms selected for reviewing the response index. The searcher reviews the response index shown in Figure 5 and creates a temporary condition such as (Automatic & Classification & Documents) & (Latent or Discriminant)*. The twenty-five highest scoring abstracts are displayed in decreasing value from left to right.

The initial letters of the browsing terms indicate the occurrence of that term in the abstract. The first column in Figure 5 indicates that the highest scoring document contains the roots: Automat, Classif, Document, Factor, and Latent.

*Abbreviated (A & C & D) & (L or D)

	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	20	1	2	3	4	5
AUTOMAT	A			A		A	A	A			A	A		A			A	A	A				A		
CLASSIF	C	C		C		C	C	C			C	C			C	C							C		
DOCUMENT	D	D			D	D	D	D	D	D	D	D	D					D								
DISCRIM							D								D									D	
STATIST			S		S				S		S		S	S						S		S	S		S
FACTOR	F		F	F					F							F								F	
LATENT	L																								
WORD			W				W	W			W			W	W	W	W							W	W

Figure 5. Response Index

The searcher scans the response index looking for combinations or partial combinations representing his desired concept. When a combination is found, the text of the abstract may be viewed on the screen by selecting the appropriate column number. Figure 5 shows: that abstract #1 is the only one meeting the specification: A & C & D & (L) and that abstract #7 is the only one meeting the specification: A & C & D & (D). As no other abstracts meet the specification, the searcher may relax his criteria on the basis of information displayed, request a new display with only three terms and view abstracts containing A & C & D.

Browsing at the display console continues by viewing the abstracts containing the three search terms, A & C & D, as shown in Figure 6. A further change in the temporary condition can be made to find additional relevant documents in which the author did not use all three terms. One could relax the condition from three terms to two terms. After viewing those abstracts the response index for the next set of twenty-five abstracts may be displayed.

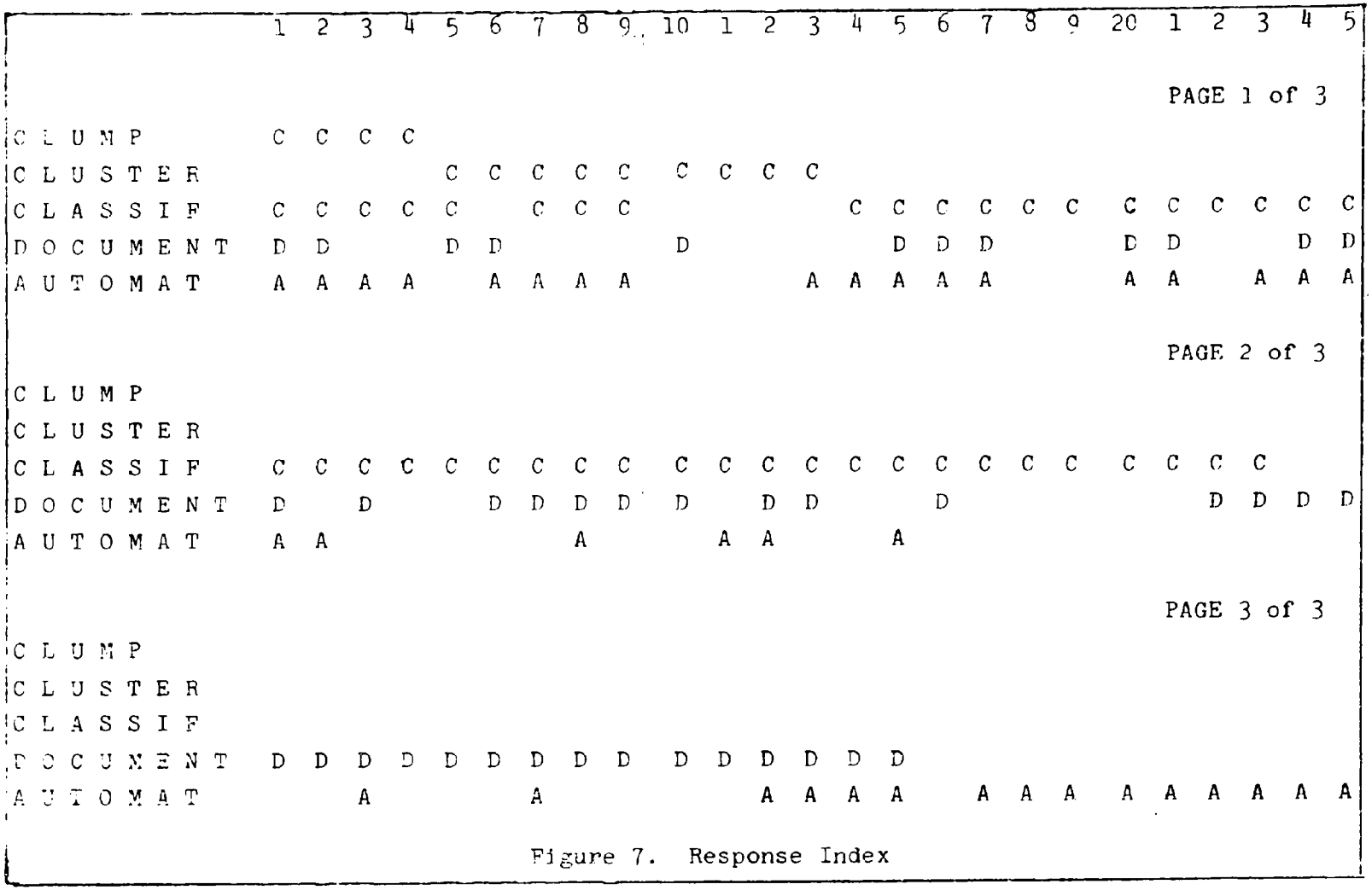
After all useful combinations of the browsing terms selected have been reviewed, the searcher may reselect another set of browsing terms. The search terms shown in Figure 4a are again displayed. Figure 7 shows the response index displayed for the new set of browsing terms: CLUMP, CLUSTER, AUTOMAT, CLASSIF, and DOCUMENT.

	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	20	1	2	3	4	5
																					PAGE 1 of 4				
AUTOMAT	A		A	A	A	A			A	A	A	A	A	A	A	A	A	A	A	A					A
CLASSIF	C	C	C	C	C	C				C	C	C	C	C	C	C		C	C	C	C	C	C	C	C
DOCUMENT	D	D		D	D	D	D	D	D		D	D				D	D	D	D		D	D	D	D	
																					PAGE 2 of 4				
AUTOMAT		A	A	A	A	A		A	A	A	A		A		A	A	A	A	A	A					
CLASSIF	C		C	C	C		C	C		C	C	C	C	C			C			C	C	C	C	C	C
DOCUMENT	D	D				D	D	D	D			D		D	D	D		D	D		D	D	D	D	D
																					PAGE 3 of 4				
AUTOMAT								A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
CLASSIF	C	C	C	C	C	C	C	C	D	C	C	C	C	C	C	C	C	C	C	C					
DOCUMENT	D	D	D	D	D	D	D															D	D	D	D
																					PAGE 4 of 4				
AUTOMAT	A	A	A	A	A	A	A	A	A	A	A														
CLASSIF																									
DOCUMENT	D	D	D	D	D	D	D	D	D	D	D														

Figure 6. Response Index

The normal sequence of documents displayed is dependent on the total score of all query terms in the abstract rather than the score of just the subset of terms. When the searcher wishes to examine all occurrences of a specific term rather than the combination of terms he may override the normal sequence and request an individual term sequence. This sequence is shown in Figure 7, which places all occurrences of CLUMP, first, then all occurrences of CLUSTER, etc. This type of display meets the requirements of exhaustive searching, as it immediately displays all references in a file of a particular term in combination with other specified terms.

Due to previous experience with keywords, descriptors, and Boolean queries, some searchers wish to emphasize or weight particular terms. The response index offers this facility by providing him a means of selecting a subset of browsing terms which are deemed most important. A complementary facility is also provided to retrieve those documents that do not contain terms deemed significant by the searcher. A listing of the highest scoring N documents based on all of the terms in the query is provided. Many relevant documents have been found on this list, which would otherwise have been overlooked, because they did not contain the previse terms the searcher thought they should.



	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	20	1	2	3	4	5
																					PAGE 1 of 3				
CLUMP	C	C	C	C																					
CLUSTER					C	C	C	C	C	C	C	C	C												
CLASSIF	C	C	C	C	C		C	C	C					C	C	C	C	C	C	C	C	C	C	C	C
DOCUMENT	D	D			D	D				D					D	D	D			D	D			D	D
AUTOMAT	A	A	A	A		A	A	A	A				A	A	A	A	A			A	A		A	A	A
																					PAGE 2 of 3				
CLUMP																									
CLUSTER																									
CLASSIF	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C		
DOCUMENT	D		D			D	D	D	D	D		D	D			D						D	D	D	D
AUTOMAT	A	A						A			A	A			A										
																					PAGE 3 of 3				
CLUMP																									
CLUSTER																									
CLASSIF																									
DOCUMENT	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D										
AUTOMAT			A				A					A	A	A	A		A	A	A	A	A	A	A	A	A

Figure 7. Response Index

In the design of the BROWSER system the decision not to record and store word position within a sentence for each term was based solely on economic reasons. Recording of word position increases storage requirements and search time. Word position could easily be added whenever the need is shown by a sufficiently large number of retrievals that fail solely due to lack of word position. By allowing the searcher to express his query in several sentences, the context of his essential concept provides additional terms that reduce the random chance of unrelated terms co-occurring in an abstract. Of the twelve abstracts containing the three search terms, Automatic, Classification, and Documents, shown in Figure 6, all of them were relevant. In other searches using the terms: Planning, Programming and Budgeting; Negative Resistance; and Cathode Followers, the results did not indicate a need for word position.

SYSTEM EVALUATION

A system evaluation was performed to identify the primary sources of error contributed either by the operating environment or by the machine algorithms. Rather than conduct the evaluation under laboratory conditions with the system designers forming queries and analyzing responses emphasis was placed on incorporating many conditions from the operational environment and including many people unfamiliar with the system to generate the search problems, to write the queries and to analyze the search response.

The processes of patent searching provide an excellent opportunity to observe the flow and loss of information through the environment into the machine and back through the environment. The process originates in the engineering department. A written description of an invention is sent to the patent department. This description becomes the problem statement for a query. A patent attorney reads and interprets the problem statement and writes a query expressing the essential inventive concepts. The query is input to the computer and the file of patent abstracts is searched. The retrieved abstracts are scanned by the patent attorney to determine which full patents should be read in detail. The patents are read, its inventive concepts determined and compared with the query submitted to the computer and to the original problem statement, for the final relevance decision.

During the evaluation procedure attorneys not only made relevance decisions but also completed forms to indicate whether the query statement contained the inventive concepts of the problem statement, and whether the patent abstract contained the inventive concepts of its patent.

A total of sixty-one problem statements were generated at three locations. Problems from New York and Hursley Park, England, were written in English; problems from Sindelfingen, Germany were written in German. All the problems were read and translated into German language queries at Sindelfingen. Twelve problems were also sent to Zurich, read and translated into German language queries. The problems represented previous IBM cases, and therefore known responses previously cited by IBM patent attorneys, and by German government patent examiners were available as a retrieval standard.

In retrieval systems, the terms used in the query by the searcher and hence input to the computer are more critical than the machine algorithm employed. Variations of queries representing the same problem were tested to determine if the system placed an undue burden on the searcher. Since many of the problems represented patent applications that IBM and other corporations had filed, the first claim of these patents was also used as a query. Thus the opportunity existed to compare the terms selected by other attorneys with those selected by IBM attorneys to represent the identical inventive concept.

The evaluation was performed on a set of 17,000 German language patent abstracts, which had been prepared by the IBM Germany Patent Documentation Group over a period of five years. Thus no special rules or conventions dictated by the system's algorithms were employed in writing the abstracts. The patent abstract file representing many subjects in IBM's area of interest provided a base to compare performance in subject areas containing a large number of similar abstracts as well as those containing few abstracts. One-half of the abstracts had been classified in Class 21, electric telecommunications, and pulse circuits of the German Federal Republic Patent Classification Schedule; one-quarter of the abstracts had been classified in Class 42, instruments and computers; one-quarter were distributed in the remaining eighty major classes. Of the sixty-one problems, thirty were concerned with Class 21, seventeen with Class 42, eight with Class 15, typewriters and printing, and six with Class 43, registers. The statistical search algorithms provided sufficient resolution to retrieve abstracts from the subject area which contained half of the file as well as retrieving from subject areas constituting a small fraction of the file.

All of the variations yielded 113 queries based on the 61 problems. The queries were searched by a computer at Gaithersburg, Maryland; the responses for 89 queries were sent to Sindelfingen, and responses for 24 queries were sent to Zurich. The output for each query consisted of the query terms used in the search, the full text of the one-hundred highest ranking abstracts, and the response index indicating each abstract that contained two or more query terms. Query terms occurring in more than one-eighth of the data base were excluded from the search by the program.

Twelve attorneys at the two locations performed a sequence of three relevance decisions, with more information available for each successive decision. The first two decisions concerned requesting additional information, with the ultimate relevance decision being made on the basis of the full patent. While scanning the response index, the first decision was whether the search terms occurring in an abstract, warranted reading the abstract. While reading the selected abstract, the second decision was whether to read the full patent. This method significantly reduced the number of full patents that had to be withdrawn from manual files and studied.

The evaluation was performed with the off-line version of the program, and without modifying the query for a second search. The

results indicated that there were no significant differences in performance due to variation of the following parameters:

(1) Problem statement originating in Hursley Park and Sindelfingen.

(2) Query written by attorney vs. query based on Patent Claim.

(3) Query written in Sindelfingen vs. Zurich.

(4) Response analyzed at Sindelfingen vs. Zurich.

(5) Searching large classes vs. small classes.

The results indicated that there were significant differences in performance due to variation of the following parameters:

(1) Problem statements originating in New York vs. Sindelfingen.

(2) Number of relevant documents found by searcher vs. time spent in analysis.

The problem statements originating in New York described a piece of equipment and hence were of a wider scope than those originating in the other locations which described a single feature. Therefore it has been learned that queries must be addressed to a single topic. Problems involving more than one topic should be translated into more than one query.

An extremely interesting statistic was found which could apply to other evaluation studies. It has been demonstrated many times that machine systems find new relevant material that had not been previously

cited. However, in this study a direct correlation was found between the amount of new relevant material found, the amount of cited relevant material found, and the time the searcher spent in reviewing the machine response. The group of searchers spending more than the median search time found 22 new relevant patents, whereas the group spending less than the median search time found only 1 new relevant patent. Further, the searchers spending more than the median time were successful in finding cited relevant patents in four times as many as the group spending less than the median time.

The three primary sources of error were: queries that did not adequately represent the statement of the problem, patent abstracts that did not adequately represent the inventive concepts of the patent, and the searcher spending insufficient time to take full advantage of the system features. The machine search algorithm which provides for natural language queries did not cause a significant number of errors.

The first two errors are not unique to this system since they constitute the heart of the information retrieval problem as well as normal human communication. With the facility in the current BROWSER system of modifying a query and re-searching it at the terminal, errors due to poor queries can be reduced. Errors due to poor abstracts can be reduced by controlling abstract quality.

SUMMARY

A text retrieval system allowing natural language queries and providing on-line browsing capabilities through an IBM 2260 display terminal has been developed and tested. The prototype system contains data bases of 25,000 German language patent abstracts, 9,000 English language patent abstracts, and 8,000 Defense Documentation Center technical abstracts.

The system was designed with a non-Boolean statement query and a ranked output to meet three types of relevance needs:

1. Any relevant document
2. All relevant documents
3. A document relevant to a portion of the query.

A Boolean statement query in serving the first need well limits its applicability to the other needs by virtue of its assigning only true or false values to the logical condition. The second need requiring an exhaustive search must allow for a ranking of imperfect documents to ensure every potentially relevant document is available to the searcher for his personal decision. Satisfying the third need requires that information on the contents of the file with respect to his query be fed-back to show him how to partition his query to find several documents each containing a different element of his problem.

Although great hope exists that the performance of man-computer retrieval systems will surpass that of the computer retrieval systems it must be remembered that a computer acting as an information channel cannot add information, at best it should not lose information. Information retrieval is no different from other human endeavors in which performance is a function of effort expended. It is interesting to note that the pioneering work in work frequency statistics was presented in Zipf's book entitled, "Human Behavior and the Principles of Least Effort."

ACKNOWLEDGMENTS

The author wishes to express appreciation to co-designer, Matthew P. Perriens; to the programming staff of Robert T. Fausey, James Lira and Eugenia N. Gregory; to John R. Shipman and Jan H. Van den Beemt, IBM International Patent Operations for their support and suggestions; and to Hans Boehmer and Franz Boehme, IBM Germany Patent Operations for their guidance and assistance.

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)	2a REPORT SECURITY CLASSIFICATION
Federal Systems Division International Business Machines Corporation Gaithersburg, Maryland 20760	UNCLASSIFIED 2b GROUP

3 REPORT TITLE

BROWSER - An Automatic Indexing On-Line Text Retrieval System

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Annual Progress Report

5 AUTHOR(S) (Last name, first name, initial)

Williams, John H., Jr.

6 REPORT DATE	7a TOTAL NO. OF PAGES	7b. NO OF REFS
September 1969	28	None

8a CONTRACT OR GRANT NO	9a ORIGINATOR'S REPORT NUMBER(S)
NONR 4456(00) b. PROJECT NO c d	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10 AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC. Other qualified users shall request copies of this report from [redacted]

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
	Information Systems Branch Office of Naval Research Dept. of the Navy, Washington, D. C.

13 ABSTRACT The development and testing of the BROWSER text retrieval system allowing a natural language query statement and providing on-line browsing capabilities through an IBM 2260 display terminal is described. The prototype system contains data bases of 25,000 German language patent abstracts, 9,000 English language patent abstracts, and 8,000 Defense Documentation Center technical abstracts.

BROWSER automatically indexes textual documents, creates an inverted file for searching the unformatted text, and creates formatted files for searching bibliographic fields. Bibliographic fields may be searched independently or in conjunction with a text search.

In addition to outputs of citations and abstract text, a new form of output - the Response Index - is provided. The response index consists of a one line entry for each abstract retrieved containing the search terms occurring within the abstract. During the browsing phase at the terminal, the response index enables the searcher to view the contents of the file with respect to his search terms and to screen the machine output for the ultimate user.

As there are no syntax, or parsing routines the search algorithms are virtually independent of language. The system has been tested on over 100 German language queries.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
On-Line Information Systems						
Information Retrieval						
Programming (computers)						
Indexes						
Automatic Indexing						
Documentation						
Subject Indexing						
Libraries						
Natural Language						
Foreign Language						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.